Technology Optimization for High Bandwidth Density Applications on 3D Interposer.

Nicolas Pantano[1,2], Cesar Roda Neve[1], Geert Van der Plas[1], Mikael Detalle[1], Marian Verhelst[2], Marc Heyns[1,2], Eric Beyne[1]
[1] imec – Kapeldreef 75, 3001 Leuven, Belgium
[2] KULeuven – ESAT-MICAS – Kasteelpark Arenberg 10, 3001, Leuven, Belgium
Nicolas.Pantano@imec.be

Abstract

3D interposers are one of just a few ways of making electronic systems faster and more powerful, but their design can be complex. This paper presents a optimization flow to assist the design of silicon interposers with the highest bandwidth density possible. Using the methodology described in this paper, simulations have shown that chip-to-chip links on a silicon interposer can achieve bandwidth densities between 250Gbps/mm and 4.5Tbps/mm depending on a wide range of parameters such as interconnect length, interlayer dielectric (ILD) material and micro-bump pitch.

1. Introduction

The microelectronics industry is currently pushing the performance of High Performance Computing (HPC) devices. Opportunities arise from increasing the bandwidth to and from the central/graphical processing unit by bringing memories and/or logic devices closer to each other, or by integrating them onto the same package [1]. These improvements can be achieved by using stacked integrated circuits (3D-SICs), where chips are stacked on top of each other and connected among them and to the outside world with through-silicon-vias (TSVs). 3D-SIC technologies allow interconnection lengths to be drastically reduced, which in turn enables a reduction in power consumption, an increase in the bitrate per channel, or both. This study focuses on a 3D-SIC Interposer technology, where active chips are stacked side-by-side and interconnected through the interposer. Compared to other 3D-SIC technologies, it shows better thermal performance, which can be critical for HPC [2]. The Interposer can be made of a silicon (Si), organic or glass substrate [3]; however, the Si interposer is preferred since a higher density of interconnects and micro-bumps can be patterned.

In this paper, a methodology to design an optimized 3D-SIC Si Interposer is presented. For a specific transmitter and receiver configuration, the maximum achievable bandwidth density is determined. This is done by identifying the optimal layer thicknesses, line width and spacing of the communication bus on the Si interposer while limiting the maximum far-end crosstalk. Finally, the bandwidth density of a chip-to-chip link as a function of the transceiver circuit parameters, the interconnection length, the use of low-k materials on the Si interposer and the micro-bump pitch scaling is assessed using this optimization flow.

The rest of this paper is structured as follows. The communication link between chips and a description of the Si interposer back-end of line (BEOL) is provided in Section 2. This setup is the basis of the interposer optimization flow developed, and is described in detail in Section 3. This flow has been used for an extensive study a complete link; a detailed analysis of the bandwidth density for different transceivers parameter configurations and considering different manufacturing options is given in Section 4. Finally, Section 5 concludes the paper.

2. Chip-to-Chip link on 3D-SIC on Si Interposer

For this study, the chip-to-chip link considered is a single-ended half-duplex link between two active chips stacked on a Si interposer. The Si interposer is made of two copper metal layers and one aluminum redistribution layer (Al RDL), similar to the one presented in [4]. The interconnection lines are in a micro-strip configuration, as shown in Figure 1, where the first metal layer is dedicated to power and ground routing, while the second metal layer is used to interconnect the active dies together. For the sake of simplicity, the RDL layer has not been represented. The BEOL is similar to that of a 65nm technology. The interconnect lines have a thickness between 500nm and 2um, a width between 350nm and 6um and their minimum spacing is 350nm.

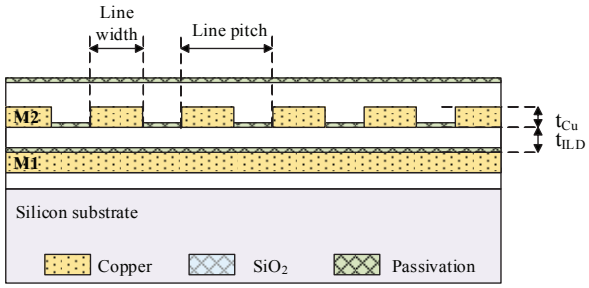

Figure 1 – Micro-strip configuration on the silicon interposer. Metal 1 (M1) is dedicated for ground and power. Metal 2 (M2) is used for signaling.

Figure 2 shows the equivalent circuit of the proposed communication link. It is composed of three lumped elements (R_s, C_{TX} and C_{RX}) and two distributed elements (r and c). The first component, R_s, is the output resistance of the transmitter and depends on its drive strength. The higher the drive strength, the lower the output resistance and the higher the current the transmitter is able to sink. The two lumped capacitances C_{TX} and C_{RX} represent the total output capacitance of the transmitter and the receiver respectively. These capacitances include the ESD protection, micro-bumps and transceiver capacitances. The last components, r and c, represent the resistance and capacitance per unit length (Δl) of the interconnection lines.

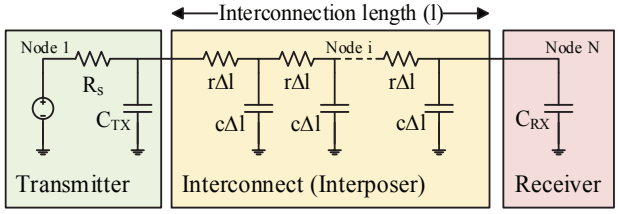

Figure 2 – Equivalent RC circuit for delay calculation of the chip-to-chip link on interposer.

3. Optimization flow

To maximize the amount of data exchanged between the stacked chips, considering a specific transmitter and receiver link configuration, the optimum dimensions of interconnection lines on the Si interposer must be found. This is equivalent to maximizing the bandwidth density, *i.e.* the ratio between the maximum achievable bitrate and the line pitch. The maximum bitrate is defined here as the inverse of the RC delay from 0% to 90% of the equivalent circuit of the link:

$$bitrate = \frac{1}{delay_{0\%\rightarrow 90\%}} \quad (1)$$

The delay is calculated from the equivalent circuit of the link shown in Figure 2. It is composed of three parts: the transmitter and receiver, represented by lumped RC components, and the coupled transmission lines, represented by a distributed RC network. If we consider each node to be discharged, a first order approximation of the Elmore delay at node N (receiver) can be obtained using the formula [5]:

$$\tau_{RX} = \sum_{i=1}^{N} C_i \sum_{j=1}^{i} R_j \quad (2)$$

where the second sum term is the path resistance, which represents the total resistance between node 1 and node i. The RC delay to change the output signal value from 0% to 90% of its final value is deduced from [5] and is given by:

$$\tau_{RX} = 2.3[R_s(C_{TX} + C_l l + C_{RX}) + R_l C_{RX} l] + R_l C_l l^2 \quad (3)$$

where R_s is the output resistance of the transmitter, C_{TX} and C_{RX} are lumped capacitances at the transmitter and receiver sides respectively, R_l and C_l are the line resistance and capacitance per unit length, l is the total length of the line and N the number of segments of the line.

The optimization flow proposed in this paper is performed in four steps:

1. For a fixed value of line pitch, interlayer dielectric (ILD) thickness, and metal thickness, we determine the maximum line width where the far-end crosstalk coefficient remains below a given limit. This coefficient is defined in [6]:

$$K_f = \frac{1}{2}\left(\frac{C_m}{C_t} - \frac{L_m}{L_s}\right) \quad (4)$$

where C_m is the sum of all mutual capacitances of one line, C_t the total capacitance of the line including the mutual capacitances, L_m the mutual inductances of one line and L_s the self-inductance of the line.

2. Within the available line width range, we determine the optimal width that maximizes the bitrate.
3. The optimal pitch that maximizes the bandwidth density is determined by repeating steps 1 and 2 for different line pitches.
4. Finally, by repeating the previous steps for different ILD and metal thicknesses, the optimal material thicknesses that maximizes the bandwidth density is obtained.

4. Link analysis

In order to perform the interposer optimization, the RLC parameters of the lines are required. These parameters are extracted from 2D field simulations (using Synopsys® Raphael™). The layer stack used in the simulations is identical to the one shown in Figure 1. The ILD is composed of silicon dioxide (SiO_2, $\epsilon_r = 3.9$) and a passivation layer mainly made of silicon nitride (SiN, $\epsilon_r = 7.0$). Regardless of the ILD thickness, the SiN passivation layer always has a fixed thickness of 50nm. To speed up the computation time, the copper layer and ILD layer thicknesses are considered equal. The copper layers have a conductivity of $\sigma_{Cu} = 5.4\times10^7\ S/m$. This value is extracted from sheet resistance measurements on the silicon interposer.

The remaining parameters required to calculate the bandwidth density are: R_s, C_{TX}, C_{RX}, K_f and the line length (l). The output impedance of the transmitter, R_s, defines the amount of current that the transmitter can sink. As an example, the strongest driver defined in the HBM2 standard has a nominal output current of 18mA, and operates with a voltage swing of 1.2 V (as defined by the standard [7], [8]), which yields an output resistance (R_s) of approximately 66Ω. As previously mentioned, the link is considered as half-duplex. Therefore, C_{TX} and C_{RX} can be considered as equal and have a default value of 200fF. Furthermore, an arbitrary but conservative value of 0.15 is considered for the far-end crosstalk coefficient, K_f.

The last parameter, l, is the interconnection length between the transmitter and the receiver. It depends on the nature of the link and the minimum pitch of the micro-bumps that connect the on-chip signals to the interposer lines. For logic-to-logic links, the physical interfaces are located on the adjacent edges of the active dies and the interconnection length can reach distances shorter than 1mm. On the other hand, the physical interface of a memory is typically located in the center of a die which increases the interconnection length [7],[8]. For HBM2 [7] or Wide-IO2 [8] memories the interconnection length can reach up to 7mm depending on the location of the logic and memory dies. Figure 3 shows a top view representation of logic-to-logic and memory-to-logic links.

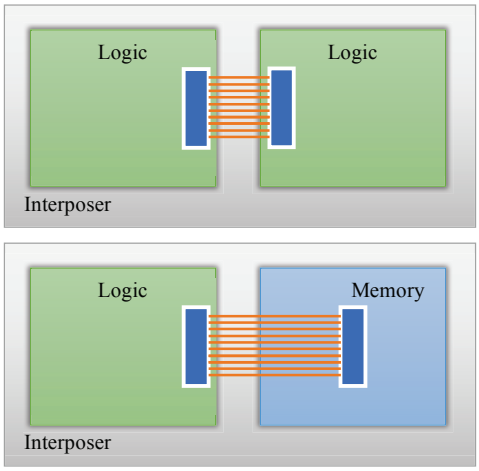

Figure 3 – Top view of a logic-to-logic link (top) and a logic-to-memory link (bottom) on interposer.

Unless otherwise mentioned, the default values of the different parameters used in the following analysis are given in Table 1.

Parameter	Default Value
Signal voltage swing	1.2V
TX output impedance, R_s	66Ω
C_{TX} and C_{RX}	200fF
Interconnect length, l	7mm
Line pitch	0.7um to 7um
Cu and ILD layer thickness	0.5, 1.0, 1.5 or 2um
Far-end crosstalk limit (K_f)	0.15

Table 1 – Default parameter values used for simulations

The analysis of the performance of a chip-to-chip link on a silicon interposer is split into three parts. In the first part, the impacts of the drive strength and the transceiver capacitances are analyzed. In the second part, the impact of interconnection length, lower-k ILD materials and micro-bump scaling (below 40um pitch) on the performance of the link is considered. In the last part, the power consumption of an optimized interposer is investigated based on post-PEX simulations for a 28nm CMOS technology.

4.1 Drive strength

The maximum bitrate achieved by a transmitter is related to the amount of current it can deliver, which is closely linked to its drive strength capability. Figure 4 shows how the bandwidth density and the bitrate of a line are affected by the transmitter's drive strength. For a 7mm line and a strong driver (66Ω, 16mA at 1.2V), a bandwidth density higher than 700Gbps/mm can be achieved with a bitrate of approximately 2.5Gbps. However, the bandwidth density drops when the output resistance of the driver increases, and for weak drivers (400Ω, 3mA at 1.2V) the bandwidth density falls to only 250Gbps/mm for a bitrate of 500Mbps per line. This shows how crucial the design of the transmitter is in high bandwidth applications.

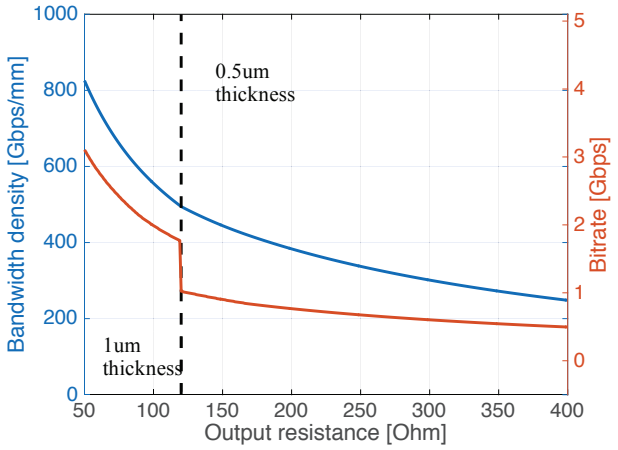

Figure 4 – Bandwidth density (blue) and bitrate per line (orange) as a function of the driver's output resistance for a 7mm line. . For drivers with an output resistance below 120Ω, the optimal line thickness is 1um. Above 120Ω, the optimal thickness is 0.5um.

Figure 4 also shows an important drop in the bitrate for transmitters with an output resistance of 120Ω. This sudden change can be explained by the way the optimization flow has been designed to achieve a maximum bandwidth density. Indeed, the optimal dimensions of an interconnect line change depending on whether the output resistance of the driver is above or below 120Ω. For drivers with an output resistance just below 120Ω (to the left of the dashed line in Figure 4), the optimal width, pitch and thickness values are of 1.15, 3.7 and 1 um respectively, whereas these dimensions become 0.7, 2.0 and 0.5 um respectively when the driver's output resistance rises just above 120Ω. This discontinuity in the dimensions of the lines has a direct effect on the delay of the line and therefore on the maximum bitrate, which consequently experiences the sudden change illustrated in the figure.

4.2 Transceiver capacitances

The parasitic capacitance of the transceivers, as well as the additional capacitance added by ESD protection circuits, is another parameter that affects the performance of the complete link and will be studied next. The impact of the micro-bumps on C_{TX} and C_{RX} are in general negligible due to their very small sizes. Indeed, typical 40um pitch micro-bumps have a self-capacitance of approximately 15fF, and this value can be reduced even further if more advanced micro-bump technologies with smaller pitches are used.

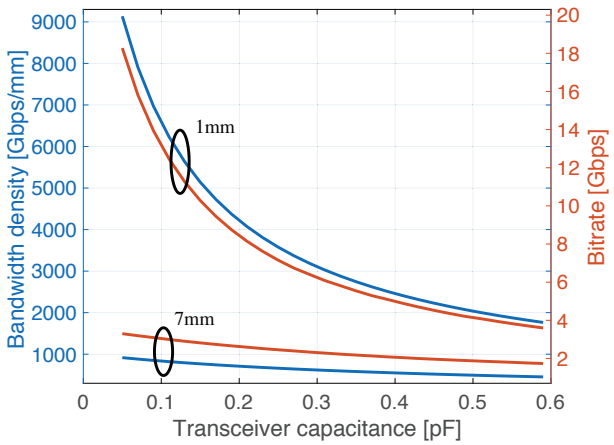

Figure 5 – Bandwidth density (blue) and bitrate per line (orange) as a function of transceiver capacitances (C_{TX} and C_{RX}) for 1 and 7mm lines with a strong driver (66Ω).

Figure 5 shows the impact of the transceiver capacitances on the bitrate for long (7mm) and short (1mm) interconnects. In the case of a long line, a quasi-linear relation between the transceiver's capacitance and the bandwidth density is observed. A transceiver with an equivalent capacitance of 50fF can yield a bandwidth density above 900Gbps/mm, which drops by half down to 450Gbps/mm when the equivalent capacitance rises up to 600fF. When the interconnection length becomes smaller, the impact of the transceiver's capacitance becomes more pronounced. In the case of a 1mm line, the bandwidth density can be as high as 9Tbps/mm if a small transceiver with a 50fF equivalent capacitance is considered, but it drops by almost 80% to 2Tbps/mm when the transceiver capacitance increases to 600fF.

4.3 Interconnection length

A critical aspect of the design of high bandwidth interfaces is related to the position of the IOs of two interconnected chips. A designer usually has the freedom to choose where the IOs are placed in logic circuits, but not in memory chips where the positions of the IOs are typically defined by standards [7], [8]. Consequently, in the case of logic-to-logic interfaces, the locations of the IO interfaces will be placed based upon bitrate requirements, whereas in the case of logic-to-memory interfaces, IOs on the logic will be generally placed closest to the memory to reduce interconnection lengths.

In Figure 6, the effect of the line length on the bandwidth density and bitrate is shown. As previously mentioned, a 7mm long line can reach a bandwidth density of 710Gbps/mm and a bitrate of approximately 2.5Gbps, but these values increase to 4.2Tbps/mm and 8Gbps respectively when the interconnection length reduces to 1mm. Indeed, shorter interconnects have a lower line capacitance and resistance, so their impact on bitrate and bandwidth density is smaller.

To further improve the bandwidth density of chip-to-chip links on a silicon interposer, two options are considered in this paper. The first one, presented next, consists in lowering the dielectric constant of the ILD material to reduce the line capacitance. The second option examines the impacts of scaling the micro-bumps in order to further reduce the interconnection length.

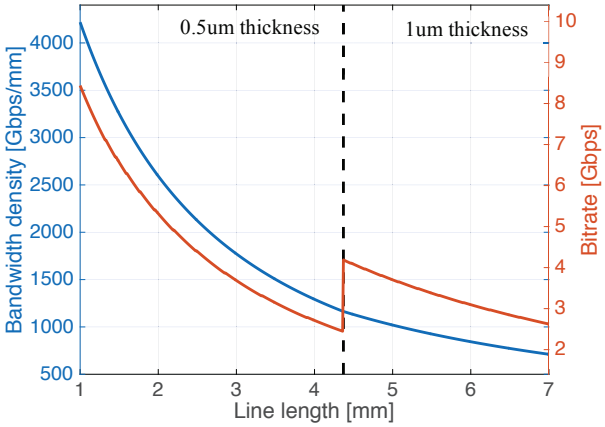

Figure 6 – Bandwidth density (blue) and bitrate per line (orange) as a function of the line length for a strong driver (66Ω). For lengths up to 4.4mm, the optimal line thickness is 0.5um. Above 4.4mm, the optimal thickness is 1um.

4.4 Low-k ILD material

One approach to scale down the delay of a link is to reduce the line capacitance by increasing the thickness of the ILD between the ground plane and the signal line, while keeping the metal thickness constant. Doing this reduces the self-capacitance of the line but increases the coupling capacitance between the lines which may limit the benefits of this approach. To reduce both self and coupling line capacitance, another solution is to change the dielectric constant of the ILD. So far, the ILD material considered on the silicon interposer is silicon dioxide, which has a relative permittivity (ϵ_r) of 3.9, but lower-k materials (ϵ_r around 3.0) can also be deposited on the interposer with similar thicknesses.

Figure 7 shows the benefits of using a lower-k ILD material. For a long line (7mm) the bandwidth density increases by 125Gbps/mm (17.5%) from 710Gbps/mm to 834Gbps/mm, and for short lines (1mm), the absolute increase is even more significant, as the bandwidth density is extended by 300Gbps/mm (7.2%) from 4.2Tbps/mm to 4.5Tbps/mm.

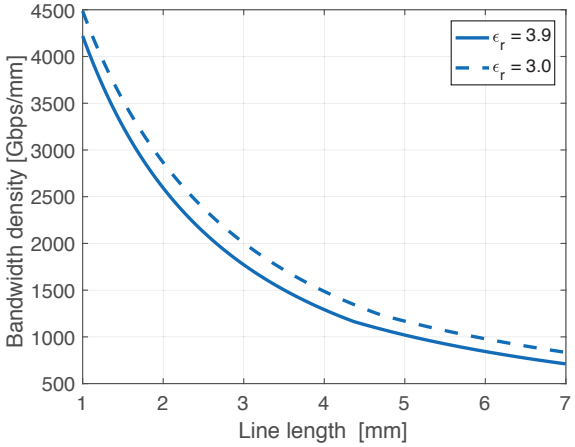

Figure 7 – Bandwidth density as a function of line length for different ILD materials (line: $\epsilon_r = 3.9$; dashed: $\epsilon_r = 3.0$).

4.5 Micro-bump pitch scaling

Another strategy to boost the performance of a chip-to-chip link on a Si interposer consists in downscaling the micro-bump pitch in order to reduce the length of interconnections. Current micro-bump technologies used for production have a pitch of 40um, but research is currently underway to develop micro-bump technologies with pitches down to 5um [9]. Scaling the micro-bump pitch leads to a substantial decrease in the area that these components occupy on a link, as shown in Figure 8. This has two major advantages. From a system point of view, as less area is occupied by micro-bumps, more IOs can be placed on a chip; but even more importantly, the decrease in micro-bump area also has an effect on the length of the interconnections between two chips. This effect, which in turn leads to an increase in the maximum achievable bitrate and bandwidth density, is explained next.

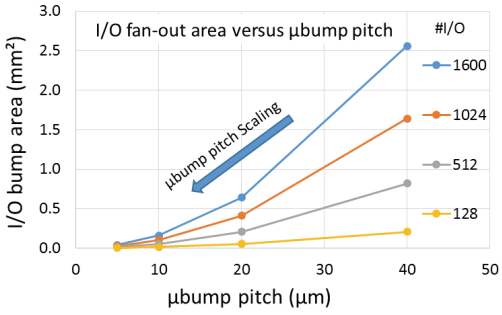

Figure 8 – Impact of micro-bump pitch scaling on the IO footprint area as a function of the number of IOs.

In this part of the study, we consider a silicon interposer with two routing layers (M2 and Al RDL, as used so far in this paper), and with IOs connected such that each line has the same length with limited fan-in and fan-out, as is the case in the diagram shown in Figure 9. Furthermore, in order to minimize interconnection length, micro-bumps are placed such as to maximize the number of rows of IOs, but this number is often limited by the dimensions of a chip or other design-related constraints. For a large number of IOs, it therefore becomes necessary to increase the number of columns of micro-bumps, but this number is also constrained. Indeed, the more columns of IOs there are, the more lines need to pass in between two micro-bumps distant from one another by the micro-bump pitch, $p_{\mu bump}$. Since each line has a certain optimum pitch, $p_{opt.\ line}$ (determined using the bandwidth density optimization flow described in Section 3), the maximum number of columns of IOs, $N_{IO\ columns}$, is equal to

$$N_{IO\ columns} = \frac{p_{\mu bump}}{p_{opt.\ line}} \qquad (5)$$

Consequently, both the dimensions of a chip and the micro-bump pitch limit the total number of IOs, but this number can be increased when micro-bump pitch scales. This is shown in Figure 9, in the case of 20 micro-bumps and a fixed routing width. When the micro-bumps have a pitch of 40um, only two rows of IOs can be used, whereas if 20um pitch micro-bumps are used instead, the number of rows can be increased to 5. This means less columns of IOs are necessary in the latter case, and consequently, the total interconnect length is reduced as explained next.

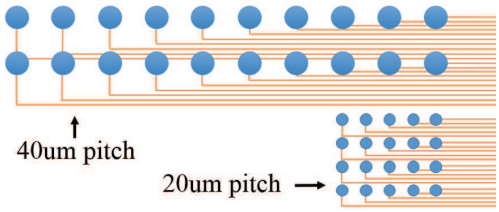

Figure 9 – Illustration of micro-bump pitch scaling for a line with $l_{min} = 7mm$. IO columns are limited to 10 and 5 for micro-bump pitch of 40 and 20um respectively.

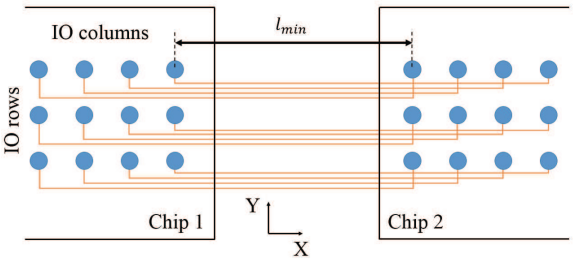

Figure 10 – Routing illustration for equal length interconnects.

Figure 10 shows that the total interconnect length between two chips is made up of two components: the distance between the two arrays of microbumps on the two chips, and an additional length of wire required to connect a line to the correct IO microbump on the chip. This can be expressed mathematically as

$$l_{total} = l_{min} + (N_{IO\ column} - 1) \times p_{\mu bump} \qquad (6)$$

where, l_{min} is the length of the portion of line between the two facing microbump arrays (see Figure 10). Equation 6 shows that the total wire length depends on both the total number of IO columns and the micro-bump pitch, which both decrease when the micro-bump pitch is reduced. Therefore, scaling the micro-bump pitch is an effective way of increasing bandwidth density and bitrate as it can lead to substantial decreases in interconnect length, as shown in Figure 11.

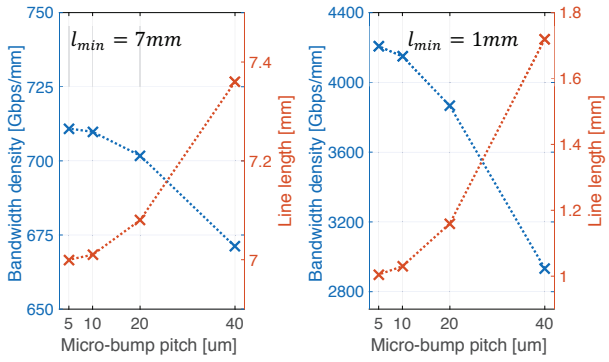

Figure 11 – Bandwidth density improvements as the micro-bump pitch is scaled for a line length (l_{min}) of 7mm (left plot) and 1mm (right plot).

The bandwidth density and the line length as a function of micro-bump pitch is shown on the left plot in Figure 11

for a 7 mm line. In this case, a 40um micro-bump pitch will limit the number of IO columns to 10, and a bandwidth density of 665Gbps/mm can be achieved, but if the pitch is scaled down to 20um, the number of IO columns is limited to 5 and the bandwidth density then rises to 701Gbps/mm. With decreasing micro-bump pitch the number of IO columns decreases so interconnection length is reduced according to (6). A smaller interconnect pitch leads to an increase in the bandwidth density, but in the case of long interconnect lines ($l_{min} = 7mm$), this improvement (36Gbps/mm) is not substantial because the relative decrease in interconnect length is small compared to the total length of the line.

However, as the minimum line length (l_{min}) decreases, the impact becomes more pronounced as seen on the right plot in Figure 11 where the effects of micro-bump pitch scaling on a shorter line ($l_{min} = 1mm$) are shown. In this case, when the pitch is reduced from 40 to 20um, the bandwidth density increases by almost 1Tbps/mm.

These observations show that scaling the micro-bump pitch in the hope of increasing the performance of a chip-to-chip link on a silicon interposer is an attractive solution preferably for short interconnection lengths. As the line length increases, the benefits of microbump pitch scaling are less evident.

4.6 Power consumption

To estimate the power consumption of an interconnect link on a Si interposer, post-PEX simulations have been performed. These simulations have shown that a 1mm interconnect line uses an energy of 0.37pJ/bit at 8.4Gbps. Alternatively, a 7mm interconnect line consumes 0.76pJ/bit at a transmission rate of 2.5Gbps. The interposer interconnect stack has been optimized using the optimization flow presented in Section 3; the transmitter is a full-swing driver operating at 1.2V, and the receiver is a chain of inverters, all designed in a 28nm CMOS technology.

5. Conclusions

This paper presents an optimization flow to efficiently design the interconnection lines of chip-to-chip links using a 3D-SIC Si interposer technology. With the help of this flow, an extensive study was performed to evaluate the performance in terms of bandwidth density of logic-to-memory and logic-to-logic links both from a circuit and manufacturing perspective.

The first part of the study shows how the performance of a link is conditioned by the design of transceivers and the nature of the link. Both the drive strength and the capacitance of the transceivers highly impact the bandwidth density of the link, especially when interconnection lengths are small. Furthermore, longs interconnects (7mm), typical of logic-to-memory links, are able to achieve bandwidth densities up to 710 Gpbs/mm, whereas short interconnects (1mm), generally used in logic-to-logic links, make it possible to reach bandwidth densities as high as 4.2Tbps/mm.

If these improvements are not sufficient to meet a designer's needs, then changes to the manufacturing process of the silicon interposer should be considered. Using low-k ILDs is one way of increasing the bandwidth density of a link; using a material with a relative dielectric constant of 3.0 instead of silicon dioxide can increase bandwidth density by 125Gbps/mm in the case of a long line, and even by up to 300Gbps/mm for shorter ones.

Another effective way of boosting bandwidth density while reducing IO footprint is to use micro-bumps with a smaller pitch, but this solution is more effective in case of shorter lines. Indeed, it was shown that smaller micro-bumps could increase the bandwidth density of short lines by as much as 1Tbps/mm – a significant improvement which makes micro-bump pitch scaling all the more worthwhile for certain applications.

References

[1] O'Connor, Mike, "Highlights of the High Bandwidth Memory (HBM) Standard," *The Memory Forum*, Minneapolis, MN, June 2014.

[2] Oprins, Herman *et al*, "Numerical comparison of the thermal performance of 3D stacking and Si interposer based packaging concepts," *IEEE 63rd Electronic Components and Technology Conference (ECTC)*, Las Vegas, NV, May. 2013, pp. 2183-2188.

[3] Hu, Dyi-Chung *et al*, "Embedded glass interposer for heterogeneous multi-chip integration," *IEEE 65th Electronic Components and Technology Conference (ECTC)*, San Diego, CA, May. 2015, pp. 314-317.

[4] Detalle, Mikael *et al*, "Interposer technology for high band width interconnect applications," *IEEE 63rd Electronic Components and Technology Conference (ECTC)*, Las Vegas, NV, May. 2013, pp. 323-328.

[5] Rabaey, J. M. *et al*, Digital integrated circuits: a design perspective, Pearson Education (Pennsylvanie, 2003), pp. 133-173.

[6] Sohn, Y-K., "Empirical Equations on Electrical Parameters of Coupled Microstrip Lines for Crosstalk Estimation in Printed Circuit Board," *IEEE Transactions on Advanced Packaging*, Vol. 24, No. 4 (2001), pp. 521-527.

[7] Jedec standard, High Bandwidth Memory (HBM) DRAM (JESD235A), JEDEC Solid State Technology Association (Arlington, 2015).

[8] Jedec standard, Wide I/O 2 (JESD229-2), JEDEC Solid State Technology Association (Arlington, 2014).

[9] Derakhshandeh, J. *et al*, "3D stacking using bump-less process for sub 10um pitch interconnects", *Electronic Components and Technology Conference (ECTC)*, Las Vegas, NV, May. 2016, pp. 128-133.